DOCUMENT RESUME

ED 069 695 TM 002 146

AUTHOR	Boldt, Robert F.
TITLE	Anchored Scaling and Equating: Old Conceptual Problems and New Methods.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RB-72-28; ETS-RDR-72-73-2
PUB DATE	Aug 72
NOTE	74p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Achievement Tests; Admission Criteria; Aptitude Tests; Competitive Selection; Equated Scores; Grade Equivalent Scales; Guidance; Personnel Selection; *Predictive Measurement; *Research Methodology; *Statistical Analysis; Technical Reports; *Test Interpretation
IDENTIFIERS	College Board Admissions Testing Program; Graduate Record Examinations; SAT; Scaling; Scholastic Aptitude Test; Vertical Equating

ABSTRACT

This paper describes several situations in which generalization of statistical results is not possible by representative sampling but which is attempted using corrections for selection of groups. The situations include hiring, admissions, differential classification, guidance, test score equating, and test score scaling. Evidence of inaccuracies of the assumptions underlying the corrections is adduced. The Pearson equations which rest on these assumptions are mentioned as a basis for scaling and equating procedures in existence. An alternative approach is suggested, and its application to anchored equating, vertical equating, scaling, and equating with mixed essay and objective material is described. The alternative approach consists of a principle for choosing objective functions whose optimization would lead to a selection of conversion constants for equating. The principle is that equal equating test scores should be associated with equal reported scores on the average. Constrained optimizations are suggested where policy considerations so indicate. (Author)

ED 069695

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

COLLEGE ENTRANCE EXAMINATION BOARD
RESEARCH AND DEVELOPMENT REPORTS
RDR-72-73, NO. 2

RESEARCH BULLETIN
RB-72-28 AUGUST 1972

Anchored Scaling and Equating: Old Conceptual Problems and New Methods

Robert F. Boldt

TM 002 146

EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY
BERKELEY, CALIFORNIA

ANCHORED SCALING AND EQUATING:

OLD CONCEPTUAL PROBLEMS AND NEW METHODS

Abstract

This paper describes several situations in which generalization of statistical results is not possible by representative sampling but which is attempted using corrections for selection of groups. The situations include hiring, admissions, differential classification, guidance, test score equating, and test score scaling. Evidence of inaccuracies of the assumptions underlying the corrections is adduced. The Pearson equations which rest on these assumptions, are mentioned as a basis of scaling and equating procedures in existence, an alternative approach is suggested, and its application to anchored equating, vertical equating, scaling, and equating with mixed essay and objective material is described. The alternative approach consists of a principle for choosing objective functions whose optimization would lead to a selection of conversion constants for equating. The principle is that equal equating test scores should be associated with equal reported scores on the average. Constrained optimizations are suggested where policy considerations so indicate.

ANCHORED SCALING AND EQUATING:

OLD CONCEPTUAL PROBLEMS AND NEW METHODS

Because of the utility of individual prediction a great many applications of psychological measurement techniques have received attention which is characterized by the relating of measurements taken under two sets of conditions. For some purposes these conditions differ in the time at which they are taken, but this difference is not the only one that offers potential utility. In this paper the focus is on situations where measurements differ not only in the time at which they are taken, but also from some necessity in the frame from which they are sampled. Consequently, generalization across the different populations is needed and this generalization goes beyond that accomplished through randomization alone.

In a hiring situation, for example, one would like to be able to predict the performance of an applicant for employment. In an academic situation one would like to be sure that an admitted candidate has a reasonable chance of completing some curriculum, or even possibly of performing well. Or one might desire, granted the foregoing goals, to select a test that might best perform these functions. And to do so, the supporting research would be very well founded on a study in which the candidates or applicants are unselectively hired or admitted depending on which context is being considered. Provided the criterion scores would be assigned on the same basis regardless of the input of talent to the school or job, the consequences of various admissions or hiring policies on the basis of test scores could be investigated. However, unselective hiring or admissions will not be allowed in practice.

More complex situations arise when many possible activities might be accomplished by people from a common pool. This situation occurs most often in the military service where basic trainees must be fractionated into a variety of types of advanced training assignments. Many of these trainees have arrived at some agreement with the service about the type of advanced training to be given, but for many trainees the assignment decision is yet to be determined when the student is in basic training. A somewhat similar situation occurs in education at the curriculum-selection stage. The "undifferentiated pool" of people from whom various specialists will eventually develop is the freshman class, or possibly the college-bound secondary school seniors, and the activities to which they route themselves are majors in the various curricula. Hence both curriculum selection and the military classification problem have in common the goal of optimum sorting of the people in a manpower pool into a set of differentiated activities, each with its own standard of excellence. In neither situation is a randomization experiment feasible.

A less obviously related situation is where one wants to supply, on a scale which is the same in some sense, scores based on different measuring instruments. In this case, the existence of the pool from which people are sorted into groups is not obvious. One might, for example, just administer both measuring instruments to the same people. But this is sometimes not possible, so Angoff (1961) describes a conceptual approach due to Tucker in which the union of the two groups is analogous to the pool from which hiring or admissions is done in the previous paragraphs. In other more extreme situations one can merely assert the existence of pools for which comparable scores are desired on certain measuring instruments when these instruments are properly scaled.

Recognizing that there is a monumental obscurity in the preceding paragraphs, the following examples are presented hopefully in clarification. The first is in that area of scaling which is sometimes called equating. Here the operational equating of the Scholastic Aptitude Test (SAT) applies. For a given administration the actual SAT may be a completely new test administered for the first time as a unit. This occurs for a variety of entirely defensible security reasons, but it does generate a problem in that the scores reported from the first administration will, in the admissions process, be compared with those reported based on performance on another SAT administered at a different time. Hence using different instruments, scores must be generated which are in some sense similar. The point of departure taken in developing the procedures currently used is that if both tests were given to the union of the two populations taking the SAT, then the scores should be so scaled as to provide the same mean and standard deviation. Both tests are not, of course, administered to the same population, but a small equating test is used instead to bridge the gap between the two populations. However, in the logic developing the equating procedures, the notion of estimating population statistics in the union of the populations definitely plays a role.

The second example which applies to supplying scores which are comparable in some sense is the achievement portions of the College Board Admissions Testing Program (ATP) or the Advanced Tests of the Graduate Record Examinations (GRE). There is an interest in having these systems of tests on scales that are comparable in some sense and the procedure which produces this comparability uses the aptitude tests associated with the achievement test systems to define similar reference populations. These are the populations

which are referred to as being "asserted" in the earlier paragraph. The reference populations for these tests are hypothetical ones in which the means, variances, and correlation of the aptitude tests are assumed as particular numbers. The actual set of values chosen for these reference population statistics are supposed to hold in a reference group for the aptitude tests and might rightly be thought of as the union of the various groups that take the achievement tests, at least in the case of the GRE. However, the referring of these reference groups to any single group of people is highly dubious from a scientific point of view and for this reason should be referred to as "reference populations, one for each achievement area," with similar moments rather than to a single reference population. In this situation the reference populations take the role of the "pool," and their existence is much more a matter of sheer assertion than in the hiring or admissions contexts.

A final example is one in which a group of people all take a common arithmetic test and then choose between a trigonometry and a business math test as the second test. It is desired to put all the scores on a common scale, and the pool in this situation is the complete group, but the math scores are to be based on the arithmetic-trig or arithmetic-business math combination and are to be put on a single scale in some sense.

It should be noted that the "similarity" of scores achieved is defined up to a certain point in the present context, but that it is also easy to misinterpret. For example, scaled and equated scores are not necessarily parallel, nor are they necessarily equivalent from a predictive point of view. None of the operations described here bear on their parallelism or predictive equivalence except rather superficially, and it may be that the desire for "similarity" which leads to an equating or a scaling is a desire

for predictive equivalence which is not achieved. That is another matter entirely and one to which this paper is not responsive.

In summary, a number of activities have been recounted that involve a pool of people to which generalization is intended and another group on which certain observations are available though not for all members of the pool. In one kind of fairly standard terminology the pool is also referred to as the "unrestricted group" and the employees are referred to as the "restricted group." This latter terminology has an employment flavor reflecting its origin. Table 1 lists decision activities and groups for the various contexts above as a possible additional source of intuitive feel for the sort of situation that is pertinent.

Logical Formulation

Like many other problems of an applied statistical sort, Karl Pearson (1903) encountered the problem of interest here and proposed a solution to it. At the turn of the century he was one of many who were tracing out the consequences of theories of evolution and natural selection. In this case Pearson was interested in the correlation between body organ sizes. Apparently Galton had earlier suggested that correlations between organ size might be a criterion for the identification of species. Pearson doubted this because he felt that these correlations would differ from local race to local race as a function of selection as much as do other descriptive statistics and hence would not be satisfactory as bases for species determination. He gave numerical illustrations including one of the influence of selecting two organs, A and B, in parents on the correlation of the like organs A* and B* in the offspring. He also developed an algebraic theory of the selective death rate. Although

Table 1

Restricted and Unrestricted Groups in

Some Personnel Decision Problems

Decision Activity	Pool (Unrestricted)	Employees (Restricted)
Hiring	Applicants	Employees
Undergraduate Admissions	Candidates	Freshman Class
Army Classification	Basic Combat Trainees	Occupational Specialty Trainees
Guidance	Freshman Class	Majors in Curricula
SAT Equating	Candidates at a Particular Administration	The Union of Candidates at a Number of Administrations
Achievement Test or GPA[a] Scaling	Achievement Test Candidates	Reference Populations

[a]Grade Point Average

these contributions seem substantively removed from the educational and psychological ones at hand, they are related in that they deal with concepts of selection in populations as abstracted in mathematical language and with the intent of reaching generalizations from population to population. For Pearson, the populations which differed did so because of natural selection as they were separated by generations. But the underlying approach, given different semantic content, has a broad range of application and constitutes the best methodology today.

In his early paper, Pearson (1908) presented his development in terms of linear regression and joint normal distributions. He considered the case of two selection variables, and in a later paper (Pearson, 1912) he relaxed the assumption that the distributions be Gaussian in character. Lawley (1943) provided a generalization of Pearson's equations to the multivariate case on both selectors and variables subject to selection. This generalization is briefly discussed in Lord and Novick (1968, Ch. 6). A variety of related problems are discussed by Federer (1963). The basic assumptions for study purposes in the present context are best presented by Gulliksen (1950) who draws out their consequences for a variety of situations. These formulae are often referred to as the range restriction formulae and are thought to be needed as the consequence of the implementation of a cutting score for personnel selection purposes.

In the present paper the notation for representing the selection of groups will be kept general. We begin by assuming that there are two populations of interest, that there are $r + s$ variables in the two populations, and that joint distributions of the variables are represented as

$$J(z_1,\ldots,z_r,\ldots,z_{r+s})$$

and

$$j(z_1,\ldots,z_r,\ldots z_{r+s}).$$

In the contexts of interest the populations corresponding to J and j are produced by a process which must be represented, and it is assumed that a reasonable representation is

$$J(z_1,\ldots,z_{r+s})\ G(z_1,\ldots,z_r) = j(z_1,\ldots,z_{r+s})\ g(z_1,\ldots,z_r)\ . \qquad (1)$$

In other words, the process produces differences between population J and population j depending entirely on the first r variables[1] and nothing else, for if some other variable were involved, then the equality would not hold. Also, one may list z variables with subscripts between zero and $r + 1$ for which the values of G and g are constant and the equality would still hold. Such variables could be reassigned subscripts larger than r or left alone. The point of this comment is that the list of variables which are arguments for G and g may exceed the list of those actually operating, without invalidating (1).

The selection process is one which requires only that the array frequencies in j are multiples of the corresponding arrays in J. Thus it is required that where a given array in J is not zero and the corresponding array in J

---

[1]While the substantive development is not at this point far enough along for substantive comment in the body of the text, it may be useful to the reader who is versed in the scaling applications discussed later to point out that the variables included in the string z_1 through z_r include reference test scores and curriculum or advanced test choice. When corrections are made during scaling procedures in the GRE or SAT, the variables of explicit selection do not include the choice of curriculum (Boldt, 1971).

is zero, then G must also be zero, and vice versa. Except for zeros the formulation of Equation (1) requires only that in arrays where J and j are nonzero, values of G and g exist which agree with Equation (1) which is certainly true since the values of the functions are real or rational numbers.

The context of the problem demands that J, J times G, j, and j times g are all density functions. Thus for J and j there exist marginal and conditional distributions, $M(z_1,\ldots,z_r)$ and $m(z_1,\ldots,z_r)$, respectively, and conditional distributions $\tilde{C}(z_{r+1},\ldots,z_{r+s}|z_1,\ldots,z_r)$ and $\tilde{c}(z_{r+1},\ldots,z_{r+s}|z_1,\ldots,z_r)$ respectively, such that

$$M(z_1,\ldots,z_r)\,\tilde{C}(z_{r+1},\ldots,z_{r+s}|z_1,\ldots,z_r) = J(z_1,\ldots,z_{r+s}) \qquad (2)$$

and

$$m(z_1,\ldots,z_r)\,\tilde{c}(z_{r+1},\ldots,z_{r+s}|z_1,\ldots,z_r) = j(z_1,\ldots,z_{r+s}) \ . \qquad (3)$$

The notation $\tilde{C}$ and $\tilde{c}$ is used to distinguish these conditional distribution functions from covariances discussed below. If Equations (2) and (3) are substituted into Equation (1) and the variable string is not included in the notation (as it will not be from here on unless needed for clarity), one obtains

$$M\tilde{C}G = m\tilde{c}g \qquad (4)$$

and if one sums or integrates over the space of z_{r+1} through z_{r+s}, one obtains the marginal distributions of z_1 through z_r in JG and jg and notes that

$$MG = mg \ . \qquad (5)$$

Finally, since Equation (5) indicates that the marginal distributions are equal, it follows that where division is possible by the marginal distributions (they are not zero),

$$\tilde{C}(z_{r+1},\ldots,z_{r+s}|z_1,\ldots,z_r) = \tilde{c}(z_{r+1},\ldots,z_{r+s}|z_1,\ldots,z_r) \quad . \qquad (6)$$

Equation (5) is a most important result and indicates that under the very general assumptions of Equation (1) and granting that the arrays involved are associated with nonzero marginal frequencies of the first r variables, the conditional arrays of the $r+1$st to $r+s$th variable conditioned on particular values of the first through rth variables are equal in the selected and unselected populations. One can therefore equate parameters of the selected and unselected conditional distributions and, knowing the parameters of the distribution of the variables of explicit selection (the arguments of G and g), make inferences about J or j, depending on to which one is arguing. Equation (6) thus rationalizes the Pearson assumptions as given, as shown in Gulliksen (1950, Eqs. 17 & 18, p. 162).

Applications

The problems given in the table all have been usually approached using the first and second moments and cross moments of the z-variables. Sometimes the problem is to estimate a variance, a correlation, a regression equation, or merely to derive a transformation on a scale. Also, all of the approaches have assumed that the regressions of z_i where i is greater than r, on the z_j's, where j goes from 1 to r, are linear. That is, if both sides of Equation (6) were multiplied by z_i and expectation taken over the space of the variables subject to selection, the result would be equations which are linear in the z_j. Hence Gulliksen (1950) takes as a point of departure that the regression equations in both the restricted and unrestricted populations are equal and linear. If we let the subscript r refer to the first r variables (z's), the ones on

which explicit selection occurs, and s refer to the variables with subscrip s from $r + 1$ to $r + s$, then the assumptions that the regression equations are equal can be written as

$$C_{rr}^{-1} C_{rs} = c_{rr}^{-1} c_{rs} \tag{7}$$

where the capital and small c's refer in this case to covariance matrices for the unrestricted and restricted populations, respectively, and the subscripts refer to the variables involved, the first subscript being for rows and the second for columns. The quantities entering into C_{rs} and c_{rs} need some elucidation. Using standard notation it would be common to refer to the covariance between z_i and z_i, for a given vector of values of the explicit selectors as $\mathrm{Cov}(z_i, z_i, | z_1, \ldots, z_r)$. The s by s matrix of such partial variances and covariances could be referred to as $C_{ss|r}$. Similarly one could imagine a $C_{rs|r}$ which would go into the computation of the C_{rs} of Equation (7). However, a little reflection will convince the reader that $C_{rs|r}$ has to equal zero since expectations averaged over the distributions in Equation (6) have to equal zero if they involve deviations of the variables subject to selection around their own array means. Hence the C_{rs} come from elsewhere, viz. the well-known least squares formula

$$C_{ee} = C_{ss} - C'_{rs} C_{rr}^{-1} C_{rs} \quad , \tag{8}$$

where C_{ee} is the matrix of partial variances and covariances of variables subject to explicit selection, averaged over the explicit selector space as distributed in the unselected population. A similar equation in small c's would refer to the matrix of partial variances of variables subject to explicit

selection, averaged over the explicit selector space as distributed in the restricted population.

One must distinguish clearly between the matrix called $C_{ss|r}$ and the one called C_{ss} in Equation (8), as one must distinguish between C_{rs} and $C_{rs|r}$ which is zero. Actually, the $C_{ss|r}$, averaged over the unrestricted space of explicit selectors equals C_{ee}, and if $C_{ss|r}$ is averaged over the restricted space of explicit selectors, the result is the matrix c_{ee} alluded to in the paragraph immediately above. Since the following assumption is commonly made in range restriction work and is explicitly given as an assumption by authors on the subject, it should be given here. That assumption is that

$$C_{ee} = c_{ee} \quad , \tag{9}$$

and it is a key one in range restriction work including scaling, equating, test selection, and validation. Using assumptions (7) and (9) as well as Equation (8), its counterpart in the restricted distribution, one has equations relating six matrices, C_{rr}, c_{rr}, C_{rs}, c_{rs}, C_{ss}, and c_{ss}. Given the observation of four of these matrices one might infer the other two. For example, it is common to assume that only C_{rs} and C_{ss} are not available for observation, but with Equations (7) and (9) they can be found.

In industrial applications one might administer a battery of tests for hiring purposes and reach hiring decisions based on the scores. Thus the test scores become the arguments of the functions named G and g. Other variables may also become of interest, such as scores on selection instruments which might replace the battery in use, and of course improved job performance is the goal of it all and that should be measured. Further, it will be assumed that

additional selection stages do not occur. Such stages could be accommodated in the logical machinery developed but it would just complicate matters needlessly. Hence we have only the restricted population, j, for which complete data are available, and an unrestricted population, J, for which data on the selection instruments only are available. Using Equations (7), (8), and (9) one can complete the variance-covariance matrix of variables as they would be observed under the theory, in the population of applicants. Using these estimated parameters one can generate the multiple correlations of interest including the ones needed for test selection. Such calculations made on the restricted population would underestimate the effectiveness of the instruments being used since their variation in the population of the hired would be restricted as compared to the unrestricted population where the personnel decisions are being made.

Undergraduate admission is very like the industrial setting, in that a set of selection instruments may be used to admit students for post-secondary school study. There are differences, however; one being that the pursuit of a high average criterion score, usually grade point average, is not as obviously desirable as high job performance is from the point of view of private industry. However, like the industrial hiring situation, the variables on which admissions decisions are made are the arguments in the G and g functions. In order to obtain the statistics necessary for computations which are routinely made in validity studies such as those produced by the College Board or the American College Testing Service, one needs either to restrict one's inference to statistics which are conditional on the variables of explicit selection, to make corrections based on Equations (7), (8), and (9) or to show that the lack of such restraint does not lead to serious bias in the specific instances

reported. To the author's knowledge, however, none of the three alternatives above are followed. Some effort probably should have been made by this time to deal with these problems, but at least one reason would make it quite difficult. That reason is that college admissions are made on a variety of scattered bases, with different reasons for different people. Test scores are not held to be the main bases of admission and quite a variety of variables are reportedly used. The variables $z_1,\ldots,z_r$, which are the arguments of the functions J and j and which are the variables on which the distributions in the very important Equation (6) are conditioned, consist of the union of all variables used in the admissions decisions for any particular group from which a college class is chosen. Use of all those variables in a validity study boggles the mind somewhat and, of course, it is never done. However, Equation (6) is the only way known to the author to argue from the groups admitted and eventually graded to the group on whom admissions decisions are made. An effective selection procedure, one making heavy use of variables which are quite valid when evaluated in the pool of applicants, would result in very low validities for these variables in the pool of selectees; these low correlations would not necessarily indicate that the selection procedure is ineffective or inappropriate. It is clear that the selectees are not the group on which to conduct studies unless the biases referred to are slight. Yet the admitted group is the group studied for obvious reasons, and the degree of bias thus introduced is not really known though one likes to say that maybe it is not large. Empirical research in this area is quite possible, though of course it does not establish applicability in general. Also, there are further theoretical problems which will be discussed later.

Army classification constitutes an additional complication of the selection problem. Here there are many criteria because for many basic combat trainees

one needs to estimate how they might do if placed in each of a variety of specialties. The pool of basic trainees must therefore be allocated to advanced training of different kinds, each trainee being taught one of a variety of skills. Brogden (1946, 1955) has dealt with this problem, and his classification theorem (Brogden, 1954) can roughly be regarded as a particularly good way of setting up the selection functions G and g. He assumes that the information by which the assignment decisions are to be made is available, and in the case of the Army system the data are indeed supplied to a central decision agency in Washington (Johnson & Sorenson, 1971). This is quite necessary as the manpower needs of the total Army need to be brought to bear on the training assignments and local decisions would constitute possibly incompatible suboptimizations. Hence the distribution of the variables used in assignment decisions is clearly known in the input population, and only performance data are needed in addition for validity studies. These data become available for those allocated to the courses, and then the equations developed can be used to infer parameters as they may apply in the input population. Thus multistage selection processes actually obtain in the Army situation; the course input is the unrestricted population with respect to a validity study where on-the-job performance is the criterion, but is a restricted population with respect to the input population.

The use of the variables of explicit selection may not be complete in the Army situation because of limitations of computer space or research time. These are practical limitations which may be overcome by the expenditure of funds or energy if one so desires. In contrast, validity studies for guidance purposes have the problem that the variables $z_1,\ldots,z_r$ are not known and indeed from the point of view of scientific philosophy may forever remain unobservable.

Those variables of explicit selection are the ones which operate to produce curriculum choice, and if they were known with any degree of confidence, our colleagues in the guidance fields would have long since told us about it. However, they may at most list certain observables which are related to, and that does not make them part of the explicit selectors, curriculum choice. To the author's knowledge the most one can do here is to try to identify variables which account for known results of selection and trust that if accurate for those variables, they are accurate for variables where the results of selection are not known. The identification of such variables, elsewhere called "surrogate selectors," is an empirical matter and has not been tested extensively to date.[2] Some accounting for selection would really seem to be needed in the guidance context because the group for whom the estimate of success is needed (counselees or at any rate freshmen) is by its definition unselected so far as curriculum goes.

Test Score Distributions Not Homoscedastic

The scaling and equating applications remain to be spelled out. They are rather different from the ones previously discussed which fit together rather as a group. It has been pointed out that there are problems in identifying and using the variables of explicit selection, such problems being quite severe in the case of validity studies for guidance purposes. However, there is yet another problem in the application of Equation (6) to these problems that has

---

[2]It has been hypothesized that if one did a discriminant function study to select those variables that best predict curriculum choice, the interest tests might prove to be most valid for this purpose. Thus the depressed validities of these tests would be accounted for, and if correction for range restriction on these variables produced accurate estimates of known covariances and then demonstrated substantially increased validities of their own, the case for the utility of interest tests would be quite strengthened. The hypothesis depends on the notion that it is interest that sorts people into curricula, more than ability.

not been mentioned. That is, it has been suggested that Equation (6) is a point of departure from which one arrives at Equations (7), (8), and (9) which are then used to work at solving the problems listed. It remains actually to get from Equation (6) to the points of departure, and the question of how it is done is at the heart of what it is that is generalized in order to provide estimates of quantities for populations that one cannot examine directly.

As mentioned above, the original discussion by Pearson (1903) of the effect of group selection was based on normal distributions, and though he relaxed his requirements in 1912, he retained the features of linearity and homoscedasticity that obtain in joint normal distributions. That is, normal distributions are often discussed as pertaining to the error of prediction in a regression equation, and it is often asserted that the variance of the errors of prediction does not depend on the values of the arguments in a regression function. Regression lines are sometimes represented as tilted straight lines around which, at each level of the predictor, are distributed values of the variable being predicted, and it is a feature of these representations that the distributions at each level of the predictor are all the same. These are examples of homoscedasticity. They probably also do not describe how test scores work--they certainly do not describe how test scores work in general.

Homoscedasticity in the notation of the present discussion is the assumption the $C_{ss|r}$ does not require the notation indicating that the covariances are dependent on the particular values of the r explicit selectors. That is

$$C_{ss|r} = C_{ee} = c_{ee} = c_{ss|r} \quad . \tag{10}$$

If Equation (10) is accepted, then the fact that the conditional covariances are averaged over the restricted or unrestricted distributions of the explicit selector to get C_{ee} and c_{ee} does not matter since either is the average of a constant which is, of course, the constant itself. With the understanding that the C_{rs} and c_{rs} are the zero-order covariances arising from calculations over the respective populations, the terms of Equations (7) and (9) are all defined and justified in terms of the model.

The position of this paper is that reliance on Equation (10) should be avoided where possible as it is detectably incorrect in a fairly regular way, at least with test scores. Application of the range restriction technique will occasionally yield results that imply impossible values of criterion such as negative scores when the criterion is a nonnegative average of ratings. It would be helpful, perhaps, to study empirically the acceptability of linearity and homoscedasticity for interesting criteria. However, the author feels safe in prejudging the outcome of such study--it is felt that lack of homoscedasticity would be apparent unless only very weak experiments were used. The wiser course of action is to find some other way to handle the problems encountered. The balance of this section will present various types of evidence bearing on assumptions of linearity and homoscedasticity in test score distributions.

As discussed earlier in this paper, the influence of the normal distribution has been tremendous in statistical theory, and in the topic at hand the first results published by Karl Pearson were couched in terms of normal theory. Though he later modified his point of view by relaxing the normality assumptions, he retained the assumptions of linearity and homoscedasticity which do obtain in the normal distribution and which, it has been pointed out, are needed for the applicability of the techniques he presented. Hence the influence of the normal

distribution is still felt. Kendall (1948, pp. 131-132) commented on the normal distribution as follows:

> The discovery that errors of observation ought, on certain plausible hypotheses, to be distributed normally led to a general belief that they were so distributed. The belief extended itself to distributions such as those of height, in which the variate-value of an individual may be regarded as the cumulation of a large number of small effects. Vestiges of this dogma are still found in textbooks.
>
> It was found in the latter half of the nineteenth century that the frequency distributions occurring in practice are rarely of the normal type and it seemed that the normal distribution was due to be discarded as a representation of natural phenomena. But as the importance of the distribution declined in the observational sphere, it grew in the theoretical, particularly in the theory of sampling. It is in fact found that many of the distributions arising in that theory are either normal or sufficiently close to normality to permit satisfactory approximations by the use of the normal distribution. Furthermore, by a fortunate accident (if one may speak of accidents in mathematics) it happens that the analytic form of the normal distribution is particularly well adapted to the requirements of sampling theory. For these and other reasons which will be amply illustrated in the sequel, the normal distribution is pre-eminent among the distributions of statistical theory.

This passage is quoted at length because it seems to the present author to put the matter particularly well--the assumption of normality leads to much interesting and suggestive theoretical work. Lord (1955), for example, used normality assumptions in connection with test score equating to derive some important and useful formulae. In fact, the handling of some problems such as those encountered in factor analysis is much easier when the assumption of joint normality is made. However, in testing as elsewhere it is not correct in practice, and while the procedures suggested through normal theory may be quite useful, their applicability arises from other considerations. In rejecting the property of normality we do not fly in the face of established expert opinion but merely follow a trend of development which has taken place in other areas. Probably, excessive adherence to the familiar procedures derived from normal theory will prove detrimental to the expeditious development of improved

statistical development and procedure unless the procedures arise from other and perhaps more realistic assumptions.

One classical way of thinking about test scores is to regard them as composed of two components, a true and an error score. Discussions of this model often refer to an error of measurement which is considered to be invariant as a function of the true score and is represented as, for example, in the College Board admissions testing program, being roughly on the order of 30 points, and this figure is rarely qualified as to different levels of performance. For example, if true and error scores were jointly normally distributed, one would expect invariance of the error of measurement with the true score level and joint normal distributions of tests. However, Mollenkopf (1949) developed a line of logic that demonstrated that unless the skewness of a test distribution were zero and its kurtosis were three, features of the normal distribution, the error of measurement would not be invariant at various score levels. Data presented by Boldt (1972) indicate that neither of these features obtain. Keats (1957) developed a different formulation of the error of measurement that agrees both with Mollenkopf's method and with Mollenkopf's data better than does the homoscedastic model (which can be seen to be clearly incorrect by examining the data Keats presented). Lord (1965) assumed that the error distribution given the true score level is a compound binomial distribution which is surely not homoscedastic. Meredith (1965) also refrained from assumptions of homoscedasticity. In fact, among test theorists it is rather widely agreed that errors of measurement are not homoscedastic over the range of true scores (Lord & Novick, 1968, p. 131).

The preceding remarks deal with errors of measurement and do not bear directly on the Pearson equations of interest, but nevertheless they very

strongly suggest that homoscedasticity between tests is not going to be realized. Lord (1965) constructed a bivariate frequency distribution from the univariate marginal distributions of two tests which purport to measure the same thing and got a rather good-looking reproduction of the actual joint frequency distribution (which nevertheless is not, according to a chi-square test, quite right). His distribution is neither linear nor homoscedastic nor would it be if it is to fit the data as well as it does.

In another context the author (Boldt, 1966, 1968) had occasion to plot array variances from bivariate frequency plots and found that the plot looked rather like a skewed but inverted cup. The skewness came from the fact that a mismatch in difficulty obtained for the tests plotted. This inverted cup has been observed in a number of other tests, both verbal and math. Figure 1 presents a typical plot displaying this inverted cup effect. What is referred to is the trend for the variances to become small at the extremes of the frequency distribution and that trend has to be appreciated in spite of a certain amount of visual static introduced by the spikey fluctuations. The latter are not represented in the polynomial distribution mentioned above, but the inverted cup effect is summarized in numerical form in Table 2. The entries for Table 2 come from scatterplots of reported SAT scores plotted against shorter tests (equating tests) that are built to the same specifications except for length. The test data presented in the first data row of Table 2 refer to the test presented in Figure 1. The first column of Table 2 gives the correlation of the variances with the associated short test score level, and the second column shows the multiple correlation of the variances with the score level and its square. Note that in some cases the gain in correlation from introducing the quadratic term is substantial, but of course the correlation with two independent

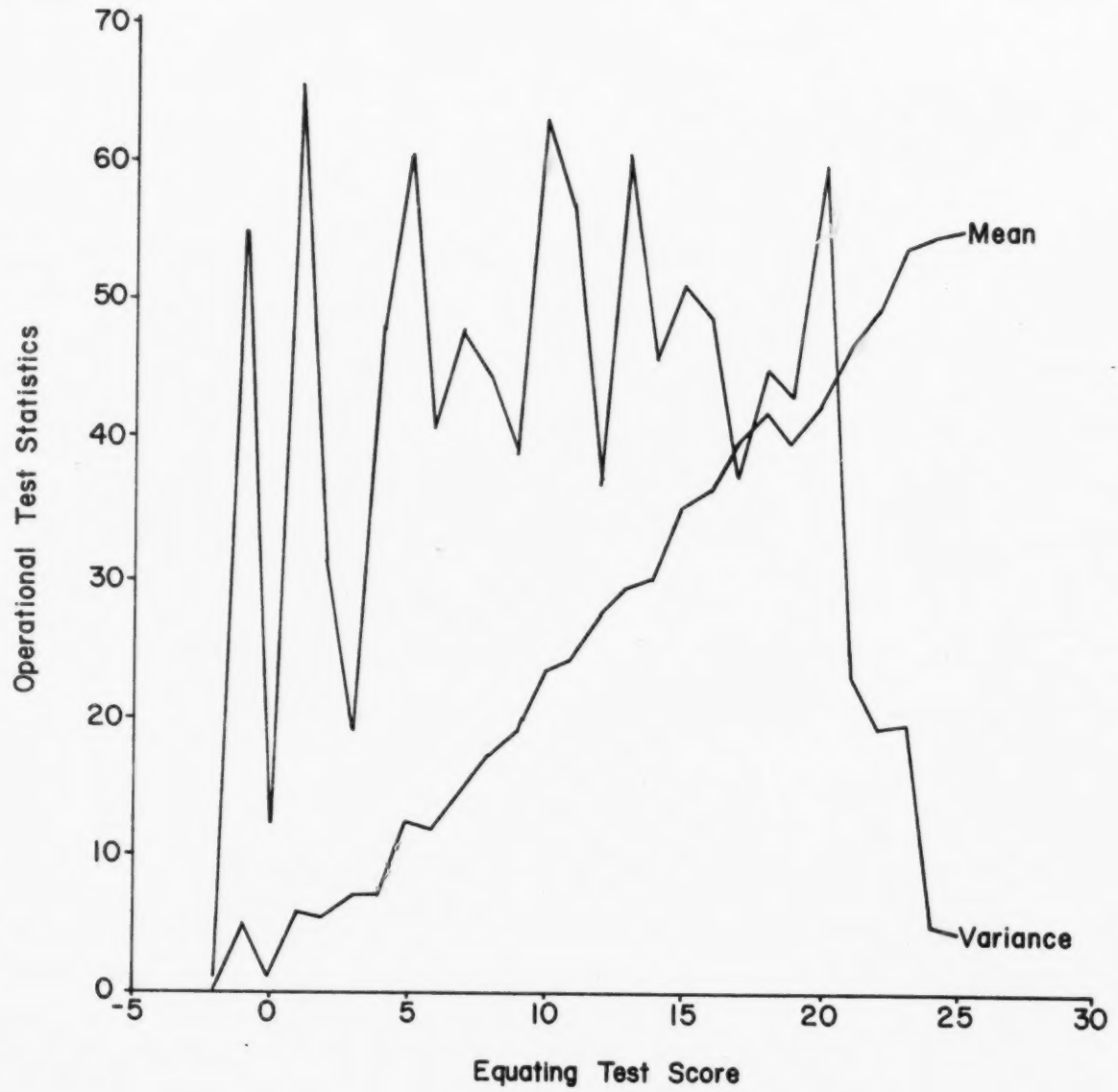
Table 2

Correlation between Linear and Quadratic Functions of Equating Test Scores and the Conditional Variances and Standard Score Regression Coefficient and Quadratic Function

Test	Correlations		Regression Terms	
	Linear	Linear and Quadratic	Linear	Quadratic
SAT-M	.3671	.6759	1.5920	-2.0396
	.0100	.8823	2.4882	-2.6494
	.3311	.8950	2.7800	-2.5862
	.3009	.5079	.8909	-1.2601
	.0158	.1810	.4856	-.5033
	.0126	.8184	2.1283	-2.2919
	.0278	.9179	2.7297	-2.8525
	.0322	.8518	2.2592	-2.3842
	.2287	.9558	3.2376	-3.5884
	.2563	.7507	2.8917	-2.7232
	.4807	.7687	1.7599	-2.3196
	.3864	.7027	1.8381	-2.3005
	.3037	.4656	1.0143	-1.3645
	.1179	.7061	2.4824	-2.6919
	.5800	.7334	1.0378	-1.6850
	.6531	.8880	1.5945	-2.3267
SAT-V	.2615	.5233	1.2869	-1.6134
	.0469	.7340	2.3549	-2.5110
	.1849	.4899	1.1059	-1.3632
	.0985	.8230	2.4892	-2.5350
	.0521	.8230	2.4892	-2.5350
	.0521	.7814	2.2954	-2.4736
	.3161	.7213	1.6617	-2.0813
	.1708	.7695	2.4980	-2.4452
	.3795	.7766	2.1826	-2.6801
	.2380	.3280	.6155	-.8828
	.4377	.7744	1.9791	-2.4998
	.4190	.7142	1.7678	-2.2619
	.2794	.4157	.8013	-1.1364
	.0534	.6771	2.5506	-2.6901
	.0978	.1020	.2072	-.1132
	.0170	.0665	.2598	-.2511

Figure 1

Operational Test Statistics as a Function of Equating Test Scores
SAT-M, N = 1196



variables is better than with one. Therefore, the data in the next two columns are of special interest. These data give the standard score regression coefficients, and the important thing to note is that in every case the pattern of signs is the same, being negative for the quadratic term. The probability that all 31 examples would have the same sign is two to the minus thirtieth, given equal odds for plus and minus, and thus the trend seems quite reliable. Hence the lack of homoscedasticity that appeared in the author's study referred to above (Boldt, 1968) is not unique to the tests used in that study but obtains between equating and operational SAT tests that are built to the same relative specifications (differing only because of length). To put these plots in the context of the Pearson equations it should be noted that the assumptions that act as a point of departure for Tucker (Angoff, 1961) can be developed from Pearson's equation, which can be clearly seen in Gulliksen's presentation (1950, Ch. 11,Eqs. 3 & 6) where the explicit selector is the equating test. Where the observations for the test being equated are based on a population which is identical to that supplying observations for the test to which the equating is being done, these equations do not require the support of derivation from the Pearson equations. But where the populations are not comparable, some derivation must be done to support his assertions. Where the populations are not quite the same, the Pearson equations are needed to support the derivation of the equating assumptions though one might conduct a study to show that the current methods are almost correct.

Figure 1 also contains a plot of SAT means for people scoring at various levels of the short equating test. Such plots might be expected to be curvilinear according to some of the results in the author's previously cited paper (Boldt, 1968), but inspection of the figures presented here shows that linearity

very nearly obtains. Linearity is also indicated by inspection of the data presented in Table 3. These data in Table 3 give the correlations between mean SAT score and the levels of the shorter test for linear, quadratic, and some cubic polynomials. Only a few of the correlations associated with cubic functions are presented since most of them suffered from near singularity of the matrix of correlations of the predictors; i.e., x^3 and x tend to be extremely highly correlated. However, the correlations associated with the linear functions are so high that further concern about linearity in this context at least seems quite unnecessary. Probably lack of linearity is greatest when the difficulties of the tests being plotted are mismatched. As a concluding note to this section of the present paper it is desired to suggest that for some purposes where a distribution assumption would be useful the following might be sufficient. Let x be a vector of random variables with mean u, let K, X, and P be positive scalars and C be a positive definite symmetric matrix of the same order as x, and u. Then let the distribution be

$$F(x) = X [K - (x-u)'C(x-u)]^P$$

if the quantity in brackets is nonnegative, zero otherwise. It will be found that if X is determined so that the function is a probability function and that K is a linear function of P, then F approaches normality as P approaches infinity. The conditional variances are quadratic functions of the variables on which the conditioning takes place and, last, the quantity P is related to the kurtosis of the univariate frequency distribution of the arguments. This kurtosis must lie in the interval from 1.8 to 3, a fact which seems to obtain for the SAT, for example. The properties of the distribution will not be pursued further here because the pursuit is not related to the following material

Table 3

Correlation between Polynomial Functions of

Equating Test Score and the Conditional Means

Test	Linear	Linear and Quadratic	Linear, Quadratic and Cubic[a]
SAT-M	.9938	.9951	
	.9875	.9957	.9967
	.9979	.9983	
	.9937	.9940	.9959
	.9910	.9972	.9985
	.9948	.9984	.9984
	.9928	.9965	
	.9933	.9966	.9975
	.9978	.9983	
	.9952	.9991	
	.9992	.9993	
	.9973	.9988	
	.9985	.9989	
	.9945	.9991	
	.9986	.9986	
	.9987	.9990	
SAT-V	.9757	.9771	
	.9832	.9840	
	.9897	.9904	
	.9961	.9966	.9967
	.9940	.9987	
	.9900	.9939	
	.9956	.9977	
	.9982	.9983	
	.9985	.9989	
	.9979	.9991	
	.9980	.9987	
	.9959	.9987	
	.9853	.9869	
	.9872	.9900	
	.9912	.9944	

[a]Missing entries arise from colinearity problems referred to in text.

(for more detail see Boldt, 1972; Press, 1972; Raiffa & Schlaifer, 1961, pp. 259-260, p. 129). However, it was felt that to suggest a possible alternative to the more usual assumptions is desirable.

Some Consequences

In terms of the activities listed in Table 1, the consequences of the foregoing section are different. For example, in the case of industrial hiring the criterion of interest is seldom a test and hence the foregoing evidence may not really apply. This is not to say that the Pearson assumptions hold in the industrial selection situation but considering the variety of criteria, the lack of standardization of criteria, and the limited size of the applicant pool, there is probably little to be concerned about in the foregoing discussion as far as industrial applications are concerned. Criticism about tests in the industrial context today arises from concerns other than those at issue in the present paper (though the present issues may bear on the feasibility of validity studies). Further, the balance of risks involved in industrial hiring are different, both for the manager and for the applicant, than in the academic context. It is felt that the technical discussion recorded here bears directly on some aspects of industrial psychology, but may not be as crucial as in some educational applications. We introduce the topic in a sense to discard it. It has served its function as part of the background of the discussion.

For undergraduate admissions the logic of correcting for selection does not seem to have been incorporated in the thinking leading to validity study designs. Certainly neither the validity study services of ACT nor the College Board explicitly incorporate corrections for selection either in estimating correlations or in correcting for selection so that regression composites can be

more accurately determined. The reasons for this may be in part practical and one may see the difficulties in the context of Flaugher and Rock (1966, 1968). Their study is one of the few in the context of college admissions that have attempted to correct for range restriction and is undoubtedly better for it. However, the corrections, being based on SAT-V and M only and correcting to the base of a hypothetical applicant pool, are highly stylized. One can learn to appreciate the formidable difficulties of accomplishing studies with better modeling processes by undertaking the intellectual exercise of trying to design a realistic selection model for the Flaugher-Rock study and then planning the implementation of such a model.

For both industrial hiring and academic admissions Novick and Thayer's (1969) study pertains. In this study range restriction corrections were made for known selection processes and the resulting corrected results compared with the known actual results. Novick and Thayer detected bias as one might expect, but the bias in the corrected multiple regression coefficient occurred when selection was quite severe, indeed, more severe than that which might be accomplished in practice on the basis of tests. It should be pointed out, however, that Novick and Thayer modeled a known and simple process and hence the rather small biases in many cases are not surprising. The situation they modeled is not that of any of the situations discussed in this paper.

Guidance practice and Army classification are rather like academic admissions and industrial hiring in that the criteria are not psychological tests and are therefore not necessarily relevant to the testing experience quoted in the previous section. In both cases the means of allocating personnel to curricula are not entirely known though in the Army situation the information available is finite and known. Also, the criteria of job performance are better standardized in the Army situation than in the industrial hiring situation and hence

empirical investigation of the criterion relationships with tests would be much more meaningful.

In all of these situations the logic of selection of group as formulated by Pearson applies if inferences are to be made about the population for whom decisions are being made. The requirements of linearity and homoscedasticity are probably not quite correct in many instances of these situations but each would profit, possibly, from separate study. Thus, for the present, one is forced to rely on the Pearson equations, hoping that they are about right and getting the best modeling of the selective processes that can be implemented realistically within constraints of budget and utility of the anticipated results. The reasoning on which some current scaling and equating processes are based is that same group selection logic of Equation (6), and the consequent equations that research or the other personnel processes rely on. However, the equating and possibly the scaling situations are rather more manageable in terms of actions that can be taken to offset the lack of validity of the Pearson assumptions, and the evidence of lack of homoscedasticity bears more directly on the scaling and equating situations, dealing as it does with the relationships of tests to tests than it does on the other processes. Alternative ways of approaching scaling and equating problems, ways that may not help much for the other processes, will be presented. These approaches will be based on linearity or on a more direct application of Equation (6) with a minimum of further elaboration of assumption. Such modifications of method might help in the scaling and equating situations because even though these procedures rest on logic similar to that of industrial hiring and the rest, the context is greatly different and the quantities required of the models are rather different.

SAT Equating

The SAT equating method that is most pertinent to the discussion here derives from equations in Gulliksen (1950, Ch. 11, Eqs. 3 & 6), which are the univariate case of the more general equations referred to earlier. Their use in this context is attributed by Angoff (1961, 1971) to Tucker. Tucker's introduction of methods at a time when Educational Testing Service did not have access to a computer and in fact accomplished test processing on accounting machine equipment is certainly not under criticism here. Very few educational technicians were well aware of the Pearson equations at that time (few of them are very well acquainted with them today) and their application to the operational processing was very appropriate and useful. The methods have been a mainstay of operational procedures for 25 years, and it has often been said that the methods are as good as can be done with the existing technology.

The methods used are to "adjust for differences in ability" and this adjustment is accomplished by treating the equating test score as an explicit selector in the Pearson equations. As has been discussed, the role of the explicit selector is that of a variable on which the selection process acts directly leaving the conditional arrays untampered with. To ascribe such a role to an equating test is recognizably a little strange since at the time the candidates sort themselves into various postures with respect to college application, the equating tests are certainly not available to act as explicit selectors. Hence the process that produces different SAT populations should not be one that is considered to produce explicit selection by equating tests. Levine (1955) introduced a modification that treats the selection process <u>as if</u> the differences in populations tested can be attributed to differences in the true score distribution alone and leaves the errors of measurement unaffected. Such an assumption is

entirely consistent with our thinking about the nature of true scores and the nature of the error of measurement. The attribution of selection solely to true scores seems to the author to have much to recommend it, and a method is presented in the Appendix for expanding the equating application of the model such that a highly efficient and automated equating procedure results. The Appendix includes a modified use of the model which constitutes a study of drift, and one can correctly infer that if the model can be used both to design and equating method and a study of a drift, it can also produce virtually drift-free equating, at least in its own terms.

The author developed the material in the Appendix about the same time as the results of the study of linearity and homoscedasticity of chance level scores became available, and the conflict between these results and the assumptions of the Levine (1955) true score equating became apparent. The true score model assumes invariance of the distribution of errors of measurement as a function of true score level, and it seemed clear that in some sense the distribution of errors of measurement must be dependent on the true score level or the results obtaining in the chance level score study would not have been found. However, returning to the results of Mollenkopf (1949) and Keats (1957), which were cited earlier, one sees that the assumption of invariance of distribution of errors simply does not hold. The wonder of it is that in the intervening decade-plus, no operational cognizance or theoretical study of the situation has been undertaken by the practitioners who use or depend on the method to produce the educational product. In defense of the Levine study it may be said that it was not basically theoretical but was an empirical demonstration that the formulae based on true score considerations have certain advantages. While it is nice for such a report to deal with theoretical problems, it is certainly

not necessary. Some subsequent not entirely verbal attempt to deal with the conceptual models and their problems is, however, long overdue.

In the material that follows, an approach to scaling and equating problems is presented which seems to the author to be reasonable and fair even given certain unrealities incurred in the use of statistical reasoning. That principle is as follows:

> Insofar as equivalence can be established, equivalent performances should receive as nearly the same score as possible.

This statement is offset not because it is particularly profound but because it states a philosophy that can be implemented almost literally in a particular situation. Obviously, the equivalent events under consideration in scaling and equating are the levels of equating test performance, and the establishment of a scoring system that yields the same score as nearly as possible is to be interpreted as a numerical optimization subject to constraints imposed in the particular context. Perhaps this is merely dodging the scientific issue, but the approach is as fair a one as the author can think of when the process that allocates people to different populations is not known or isn't adequately modeled. Implementation of such a policy is an explicit attempt at fairness, if not of rigor.

It should be pointed out that while the approach introduced above avoids explicit assumptions of linearity and homoscedasticity, the assumption that explicit selection is based on the equating test score is not one that can be avoided when making theoretical statements about the equating. Clearly, the equivalent events must be observable if the approach is to be applicable in the absence of a statistical model, and the events which are on hand to be used as

observables in more than one population are the scores achieved on the equating tests. If the population selection process is to be represented at all, the author has not been able to get back of Equation (6).

It was indicated above that minimizations are to be involved, perhaps constrained minimizations. These will not be traced out in this text since the derivation of optimization procedures is another matter and need not be of concern here. The optimizations involved may not have been feasible in many cases in 1946 when the Pearson equations were adopted, at least not feasible within the context of the testing operation being discussed, but they are quite feasible now with the current equipment. They are not optimizations that are difficult within the state of the art--complications would arise mainly from start-up costs due to rearrangements of data logistics and programming.

The first equating method to be presented treats the test scores merely as indices, and they do not enter directly into the calculation except as indications of groups to which scores belong. The method assigns a score, called a reported score, to each level of the operational test score. For an "old" test it is assumed that a mapping of operational test scores y into reported scores S_y had been achieved and that mapping is taken for granted. A new test z is operational, has been administered, and a mapping of levels of z into reported scores s_z is the equating desired. An equating test x has been administered to both the old and the new populations. If n_{xy} and n_{xz} are the frequencies of cases receiving scores of both x and y, or x and z, they are clearly observable, and a quantity V_x can be defined such that

$$V_x = \sum_y \left[\frac{M_{xy}}{\left(\sum_y n_{xy}\right)} \right] S_y$$

is the average reported score assigned to equating test score level x. As nearly as possible, we would like the V_x to hold in the new population. The following least squares minimization would accomplish that:

$$\tilde{M} = \sum_x \left(\sum_z N_{xz}\right) \left[\sum_z \left(\frac{N_{xz}}{\sum_z N_{xz}} \right) S_z - V_x \right]^2 . \qquad (11)$$

Thus one goes from the policy directly to the function to be minimized for equating purposes. Unfortunately, there are infinitely many sets of values s_z that reduce the value of $\tilde{M}$ in Equation (11) to exactly zero. That not all of these would be acceptable as equatings can be seen because the number of levels of the equating test for which the values V_x are given are less than the number of levels of the operational tests, at least for tests of the type under discussion. If additional equating tests were used adding the constraints of other administrations, and if enough such tests were used, the expanded version of Equation (11) would eventually reflect overdetermination such as is required. For the SAT, <u>at least</u> three such tests would be needed.

To bring the equating problem within reach, it seems easier to reduce the number of quantities to be estimated. Note that in the formulation above the scores are used only as indices. In fact, the reported scores do not have to be monotonic with operational test scores as the equations above are written, though they probably would be since V_x would be monotonic with the equating test score from earlier equatings. One may substitute for s_z in Equation (11) the quantity $Az + B$ so that the reported score is simply a linear function of the formula score. This does not assume directly that some linearity obtains. It simply limits the range of acceptable transformations to those which are linear. Thus the dozens of values that must be estimated in the

equating are limited to two--A and B . Probably a better choice of transformations is possible, but when such are developed, these will probably come from item characteristic curve theory or a strong true-score theory such as that due to Lord (1965) and cited earlier.

One might well wonder where Equation (6) which was so carefully developed has entered into the equating method mentioned here. The answer is that Equation (6) enters in when one makes the claim that it makes no difference which group was used for the equating. Equation (6) says that the quantities

$\left(\frac{n_{xz}}{\sum_z n_{xz}}\right)$ and $\left(\frac{n_{xy}}{\sum_y n_{xy}}\right)$ would be the same in either the old or the new population

within sampling error and up to a constant of proportionality obtaining in case the sample sizes are different. The policy from which Equation (11) stems may be sufficiently acceptable for the method to stand alone, and if we further accept the range restriction logic which is currently explicitly used in operational methods, we may more rigorously generalize the application of the results. However, we do not develop the method based on the applicability of the range restriction assumptions.

In the paragraphs immediately above a general expression, $\tilde{M}$, has been developed as an objective function whose minimization would lead to an equating of a new test Z . The role of the range restriction equations in interpreting the results of such an equating is pointed out, and it is also pointed out that by expressing the s_z as a linear function of the formula scores, the number of parameters to be evaluated in the optimization of $\tilde{M}$ is small enough that the determination is made with some reliability. However, the restriction of the s_z to linear transformations of the test scores is too limiting because

policies concerning the assignment of scores at the extremes of the test score ranges override by fiat policies aimed at producing a stable equating. Thus there is a policy that says that a score which is at the chance level (a negative formula score) should be assigned a reported score not to be in excess of some predetermined number.[3] Hence, we do not merely modify the expression of Equation (11) by writing s_z as a function of A, B, and the test score of the succeeding paragraph and call it an equating method because it doesn't handle the end-point problem. Nor does it take into account double- or multi-part score equating. What is done to finish the equating method is to write the objective function as depending on several equating tests and append it with constraints which for policy reasons are desirable. The resulting optimization will be in the form of a non-linear program, possibly, or some other numerical method which is not of interest here so long as the optimization is feasible in the practical sense.

The term objective function, as used above, refers to that mathematical expression whose optimization leads to quantities required in the equating. The quantities required are the scores to be reported, s's, for each category of performance (formula score) on the operational test. However, the particular objective function, $\tilde{M}$, given in (11), does not contain the properties needed, and it is desired to generalize it to include multi-part score equating. The term multi-part, instead of double-part, is used because the transition to more general notation is as easy as to restrict the notation to two tests. This can be done in a strictly formal way by including a subscript i to the x and generalizing (11) as follows:

---

[3]The author's suggestion of a method to deal with this policy should not be taken to suggest his concurrence with it. The policy is used because tests whose difficulties are improperly pitched scale out so that chance scores are misleadingly high. Test construction according to proper test specifications is a more appropriate remedy to the problem.

$$\hat{M} = \sum_i \sum_{x_i} (\Sigma n_{x_i z}) \left\{ \sum_z \left[\frac{M_{x_i z}}{(\Sigma_z n_{x_i z})} \right] S_z - V_{x_i} \right\}^2 . \qquad (12)$$

The formula for $\hat{M}$ is put into the multi-part score equating context by defining the terms x_i and $n_{x_i z}$. We let i refer to a particular equating test-operational test combination and x_i refer to the levels of equating test performance in the ith combination. Then $n_{x_i z}$ is the number of cases (in the current operational administration) that scored at the x_ith score level on the equating test that pertains to the combination i. Finally, V_{x_i} is calculated as indicated in the text immediately preceding (11) with the exception that equating test levels indicated by x in that expression and the operational test y are the equating test-operational test combination associated with the subscript i in the formulation of Equation (12). That is, V_{x_i} is the observed average reported score at the score level x where the score level is that of the equating test associated with the equating test that goes with the subscript i. These associations are routinely made in choosing the equating tests that are to be used for operational administrations and have been last studied by McGee (1961).

With the definitions above, and the restraint that the

$$s_z = Az + B ,$$

we arrive at the following objective function.

$$M_1 = \sum_i \sum_{x_i} (\Sigma_z n_{x_i z}) \left[\frac{\Sigma_z n_{x_i z}(Az + B)}{\Sigma_z n_{x_i z}} - V_{x_i} \right]^2 \qquad (13)$$

Thus, all the values of (13) will be known at equating time except A and B. The choice of A and B such that M_1 is at a minimum is the means of obtaining conversion parameters--the function (13) has a unique minimum in which the formulae for A and B are expressible as the solution to a pair of simultaneous linear equations.

It is not the intention of this author to display the formulae that minimize M_1 since the computational steps are not relevant to this discussion. The intention of this discussion is to give the equating rationale and from it develop an objective function that accomplishes the task set out by the rationale. Equation (13) gives an objective function that does this in at least as good a sense as the current methods, in that it develops a single set of conversion parameters. It is also similar to the present methods in that it uses Equation (6) to arrive at the justification for using V_i as the desired average of reported scores for a given equating test score, assuming that the difference between the population for which the equating is being done differs from that being used as part of the equating process by explicit selection on the associated equating test. The method of Equation (13) is somewhat superior to the current methods, in that it combines double-part score equating into a single optimization where the choice of conversion parameters is determined in reaching an optimum compromise between the several part score equatings. However, the conversion parameters obtained by minimizing M_1 might very well not be acceptable in the case of a test such as Hebrew Achievement. In this case the conversion parameters might indicate that a score at the chance level would be converted into a reported score of 400, say, which seems like too much for a test performance at the guessing level.[4] When such a set of conversion parameters is obtained

[4] In footnote 3 it was commented that appropriate test specifications should be established and met. Here is an example--the test is too hard for the examinees.

under existing procedures, a set of more or less arbitrary steps are applied to yield a point-to-point equating or a second linear transformation is chosen produce a reported score through part of the range of z. If such departures from a single linear conversion are to be gotten by minimizing one of the forms of M, clearly it cannot be that of Equation (13) but rather that of (12) together with a set of constraints. The development follows.

First, note that two policies are involved. One is that the transformation from formula score to reported score will be linear; the other is that scores in the chance level will not exceed some specified amount. When these two notions conflict is when the linear transformation produces a line that never reaches down to the desired score. The situation is pictured in Figure 2. In Figure 2 the formula score of zero is the coordinate of point A which is by the policy to have a reported score of not more than 400. Note that the unconstrained linear solution would not satisfy this policy and hence at some point, indicated by the abscissa of x on the figure, will be the lowest point at which the linear solution applies. Then as one moves down the formula-score scale to the point zero, one wants the reported scores to change regularly and smoothly to the point where the chance score has a reported score as near as possible to the linear solution and still within the range admissible by the policy. This would be the maximum acceptable score at the chance range, and it is taken as 400 for the purpose of discussion. Then as the formula scores move further into the negative range, one need only require that the reported scores are nonincreasing. Assuming that changes in the constraints on extreme cases will not noticeably affect the linear solution, Figure 2 might also be taken as a picture of the constrained solution with a relatively systematic trend from the linear solution to the chance level score of 400 with nonincreasing scores shown as one observes reported

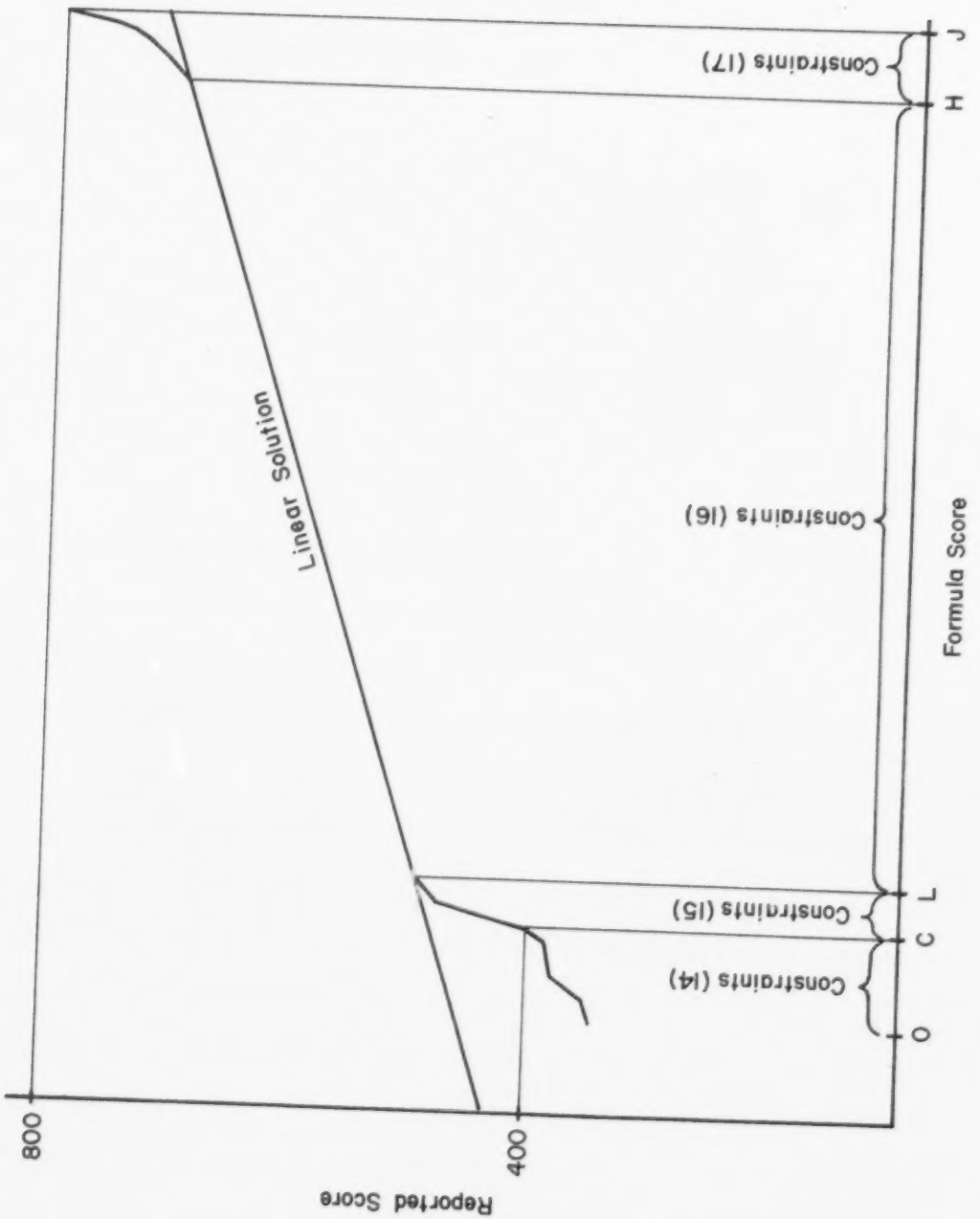


Figure 2
Example of End-Point Adjustment for
Transforming Formula Scores to Reported Scores

scores corresponding to successively decreasing scores below the chance level. One must, to fit this into a satisfactory equating scheme, express these constraints in mathematical terms so that they can become a part of the optimization that is a part of the equating scheme.

To arrive at a mathematical statement of the constraints the use of z as a subscript must be discontinued as it will be necessary to represent the "first" (lowest), "second" (next lowest), etc. values of z; the z's are not equally spaced if they are formula scores. Hence z_j will be the jth value of z in order where z_0 is the smallest, and s_j will be the reported score associated with that formula score. Suppose further that the lowest score to which the linear conversion is to apply is the Lth score, and that at the chance level $j = C$, so that $z_C = 0$. We then have three types of constraints on the s_j, those from $j = 0$ to $j = C - 1$, and those from $j = L$ and up (we are assuming that $s_C = 400$). Constraints for the values of s are

$$s_j \leq s_{j+1} \qquad j = 0,\ldots,C-1 \qquad (14)$$

$$2s_j \geq s_{j+1} + s_{j-1}, \text{ and} \qquad j = C+1,\ldots,L \qquad (15)$$

$$s_j = Az_j + B. \qquad j = L,\ldots,H \qquad (16)$$

The constraints (14) assure that the reported scores in the chance range increase monotonically with increasing formula scores. The constraints (15) assure that the reported scores which progressively move from the straight line conversion to the policy implied at a formula score of zero form a curve which is convex upward and to the left as in Figure 2. The constraints (16) assure that the conversion is linear from the point where $j = L$ up to where another policy might take over (to be discussed below). Note that due to the defined limits where

constraints (15) and (16) apply, both forms of the constraints apply to the points where $j = L$ and $j = L + 1$. This is done so that points fitted under the constraints (15) can never wander above the extension of the linear converted line--a condition which conceivably might obtain otherwise but which won't with the constraints written with limits as set above.

Before rewriting the objective function further, constraints may be needed. Unlike Hebrew or Math Level II, some tests may not scale out to a reported score of 800. This condition also is shown in the upper right of Figure 2. One might be inclined to feel that to be fair to the candidates it should be possible to achieve a score of 800 on any test[5]--clearly with the linear conversion shown in Figure 2 such will not be the case. Therefore, it could be desirable to depart from the linear conversion in the manner shown in the figure; that is, smoothly moving to a maximum of 800 for s_J with the last point on the line of linear conversion at $j = H$. This is accomplished through the addition of the constraints

$$2s_j \leq s_{j+1} + s_{j-1} . \qquad j = H,\ldots,J - 1 \qquad (17)$$

Note that the inequality of (17) is the reverse of the inequality of (15) producing the concavity and convexity, respectively. Again, the overlap of points involved in the inequalities (17) and those of (16) assure that the solution when plotted will not reveal that the reported scores are less than the linear conversion extended.

---

[5]An enforced scaling out to 800, in the case of the College Board tests, occurs because the possibility should exist for any candidate to achieve the maximum score, whatever the form he is administered. Eight hundred is supposed to be at the top of the expected range of difficulty of most of the tests in the College Board Admissions Testing Program and is therefore used as a representative top. Actually, the "fairness" achieved by such enforced scalings is mainly illusory, and true fairness is accomplished by establishing appropriate statistical specifications and then meeting them.

With the symbolic statement of constraints that provide reasonable conditions for solutions to the end-point problems, the equating method suggested here is to find s_j's which minimize $\hat{M}$ (Equation (12)) subject to constraints given in (14), (15), (16), and (17). While the procedures for finding the s_j's, that is, the reported scores, exist, they are not to be displayed here. The computational problem is one of nonlinear programming to which approaches are discussed by Fiacco and McCormick (1968). What is displayed here is the goal of the procedure; that is, the function to be optimized together with the constraints. In this procedure the assumption that would be associated with the selection Equation (6) is that the present population differs from the respective ones entering into the multi-part score equating by virtue of explicit selection on the equating test used for the particular part, or at least the selection acts this way. Particularly, the consequence of this assumption is that the n's of Equation (12) are proportional whether they are observed in the population to which the test being equated to was given or in the population to which the test being equated was given.

In implementing the equating method dsscribed here, assuming the n's and V's are available, one would first fit Equation (12) using only the constraints (16), setting $L = 0$ and $H = J$. That is, fit a straight line and see if it works. If it goes high enough and low enough, the equating is finished and score conversion can commence. If not, and if the problem is at the low end of the scale, refit (12) but under the constraints (14), (15), and (16). If not, and the problem is at the high end of the scale, refit (12) but under the constraints (16) and (17). Problems at both ends imply refitting with all constraints, (14) through (17), imposed. In this way, unless a definite problem is observed, the conversion will be linear and probably this will be the case most of the time. When the constraints must be invoked, the

technique of fitting (12) though with the constraints imposed produces the best equating one can get within those constraints. If the equating produced in this way is not satisfactory, it is because some condition is not included in the constraints or because some redefinition of the objective function (12) is needed.

It should be pointed out that when constraints other than those of (16) are used, one must decide where in the formula-score scale one will depart from the linear conversion. One intuitively reasonable way to make this choice is to pick values of L and H such that the average departures from the linear conversion are acceptably small. This average departure will probably be overestimated by calculating the average as if the departure from linearity were made using another straight line. The overestimation is expected because of the convexity and concavity due to constraints (15) and (17) which result in smaller deviations toward the more dense parts of the distribution of formula scores. Thus as more and more formula scores are involved in nonlinear constraints, the average deviation from linearity will decrease, the total number of people off the line will increase, and the achievable minimum of the objective function will probably increase. At the present time the author knows of no obviously correct way to trade-off these various effects, and it will probably be useful to examine some particularly troublesome College Board Achievement areas such as Hebrew, German, and Mathematics to see what the results of imposing various constraints might have been. In this way a satisfactory rule, probably in the form of choosing a fixed value of H or L if nonlinear constraints are to be used for a subject matter area (or possibly for all), can be formulated for some time to come.

The situation described in Figure 2 and for which constraints are introduced above do not exhaust the problems that may arise regarding end points. Another

kind of problem is, in a sense, the opposite of the difficulty previously described. That is, instead of having a situation where the conversion line doesn't reach down to 200 or up to 800, one may have a situation where a linear equating would scale well up over 800 and/or well below 200. Since, by policy, scores are not reported above the former or below the latter the result would be a pileup of scores at the extremes with the consequent loss of discrimination. To provide discrimination at the upper end it may be desirable to introduce a bend in the scale at some point below that at which the linear conversion would imply a score of 800. Suppose that $\tilde{H}$ is the smallest value of j for which the linear conversion is greater than 750. Then the constraints

$$2s_j \geq s_{j+1} + s_{j-1} \qquad j = \tilde{H} - 1, \ldots, J - 1$$

$$s_J = 800 \qquad j = 1, \ldots, \hat{H} + 1$$

would begin the bend at 750 and scale up to 800. At the lower end, if the pileup were observed, the constraints

$$2s_j \leq s_{j+1} + s_{j-1}$$

$$s_0 = 200$$

where $\hat{H}$ is the largest value for j for which the linear conversion is less than 225 would begin the bend just before the scale reaches 225 and would scale down to 200. Clearly the choice of 225 and 775 for the points of introducing the bend are arbitrary but it is hoped that a minimum of disturbance of the linear fit would be introduced while still attaining some kind of discrimination at the extremes.

Achievement Test Equating

The reader will notice that the previous section is headed SAT Equating, but where the end-point troubles arise the achievement tests have been referred to. This is, of course, because the procedures that might be used for SAT equating should have a provision for handling tests that somehow do not behave statistically as one would normally expect. The most striking examples of this type of test behavior actually come from the achievement test areas and, hence, have been mentioned as the examples. Whether the methods suggested actually apply to these tests is another question. The author's position is that they clearly do as long as the equating test is a miniature of the operational test. When it is not, but has another subject matter, the rationale for using the method would differ somewhat, and it may be somewhat different in form.

For most subject matter areas, the difference between SAT equating and Achievement test equating is in whether the equating test is part of the operating test. In SAT equating, the equating test contains different items and is timed separately; in Achievement test equating the equating test is actually part of the operational test. While these differences have a great deal of impact on operational procedures during test construction and administration, the resulting scores fit the Pearson-type formulation in that the equating test and operational test are jointly distributed variables whether one is embedded in the other or not, and explicit selection is assumed on the equating test and can be represented by taking products of functions as in Equation (1).[6] Homoscedasticity is rejected even though the equating test is

---

[6] In this comment the existence of context effects and the lack of independence due to enforcement of a single timing on both the equating material and the other material are not considered. Actually, the proper form for the operational equating experiment is one where separate timing of the anchor test occurs if one is to be used. However, the representation of these context and timing effects will not be presented in the symbolism used here.

embedded in the operational test, the array variation being supplied by that portion of the operational test which is not part of the equating test. It is as reasonable to accept the policy that equal equating-test performance should imply equal reported scores on the average, possibly even more reasonable since the equating test is, in the case of achievement tests, a part of the actual operational test performance. Also, the end-point considerations apply in the same fashion whether the tests be for achievement tests or aptitude tests, or at least the policy as understood by the author does not include any substantive differentiation. Hence no reason is apparent why the procedure should not apply to achievement tests as well.

This question is raised because in one other equating model, that consistent with Levine (1955), the procedures for internal and external equating tests are different. This is when the equating test is assumed to have a true score component which is equal to that of the operational test up to a linear transformation. The difference in whether the internal or external equating test is used comes in whether or not the errors of measurement of the equating and operational test can be assumed to be uncorrelated. If the equating test is external the errors of measurement are assumed to be independent of those of the operational test; whereas if the equating test is internal, a part-whole correlation between errors of measurement obtains. This difference leads to different interpretation of computational results and hence to different computational procedures. Therefore, under the traditional true-score model assuming explicit selection is on the true score, internal and external equating must be treated differently.

However, we have rejected the traditional true-score model. It is conceivable that true and error scores are uncorrelated, but it seems almost certain considering the evidence adduced earlier that they are not independent and that the range restriction equations will not work. This is too bad, because the idea

of assuming that range restriction occurs explicitly on the true score is highly appealing and would allow one to build an equating policy that "on the average equal true scores imply equal reported scores," if the correct (or at least a reasonable) true-score model were available that afforded the development of feasible numerical procedures, but at this point the author does not know how to do such equating. Rather we are not using any true-score logic, and the procedures suggested are based on notions by which we are unable to develop a distinction between internal and external equating tests.

Vertical Equating

The rule that equivalent events should be assigned reported scores as similar as possible suggests an approach to the very difficult problem of vertical equating. The present methods are, in the view of one of their originators (Lord, 1969), quite unsatisfactory.[7] This comment was not meant to be critical, of course, but expresses the difficulty of the problem attacked. In the same communication he urged an approach to vertical equating through the use of the Rasch model, and it is assumed that other latent trait models would do as well. However, because it fits the context of designing equivalent scores for equivalent performances, another approach is outlined below.

The vertical equating problem arises when, for example, one wants to develop a series of examinations and a scoring procedure that would allow one to trace the development of a skill. Suppose one would want to trace the development of arithmetical skill from second to sixth grades, doing this with a series of tests of appropriate difficulty. Perhaps there are three tests to be given, an easy one, a middle one, and a hard one. To equate them one might administer the easy

---

[7] He adds, "I do not say this in criticism, since I was partially responsible for the method...."

test and the middle test to the third grade, and administer the middle test and the hard test to the fifth grade. The particular pattern of administration is not at issue here as long as some pattern of administration is accomplished that relates the tests by administering combinations to various groups of people. This will be commented on later.

The equating would in this situation be accomplished by applying the principle that a person should, as nearly as possible, get the same score, no matter which test he takes. For example, if a third grader gets a score of u on the easy test and a score of v on the middle test, this is taken as evidence that the scaled score associated with a test score of u on the easy test should be similar in value to the scaled score associated with a test score of v on the middle test. This principle can be implemented as follows. Let

i be a subscript for individual;

j, j' be subscripts for test;

k, k' refer to a score level within test (k ranges from one for the lowest score to K_j where K_j is the number of score levels in the jth test);

δ_{ijk} be one if individual i scored at the kth score level on the jth test, zero otherwise; and

S_{jk} be the scaled score associated with the kth score level on the jth test.

The equating problem is to find the S's given that the δ's have been observed. These S's are chosen so as to minimize $\tilde{\theta}$ where

$$\tilde{\theta} = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{j'=1}^{J} \sum_{k=1}^{K_j} \sum_{k'=1}^{K_{j'}} \delta_{ij'k'} \delta_{ijk} (S_{jk} - S_{j'k'})^2 ,$$

where i ranges over all people in the equating experiment, and J is the number of tests which was by hypothesis three in the example. Note that a minimum value of $\tilde{\theta}$ equal to zero can be obtained if all S's are taken as zero. This is, of course, a nonuseful solution to be avoided by constraining the solution such that the sums of squares of S's , or possibly the sums of squares of S's for a particular test, be set equal to an arbitrary positive constant using a LaGrange constraint. The choice of the constant would be for convenience. Another constraint would be needed to establish a zero for the scale. Clearly, the value of $\tilde{\theta}$ is invariant under additive shifting of the S's , so probably it would be convenient to set some arbitrary average of the S's equal to zero or a constant. Once a solution is found in terms of the arbitrary constants, a linear translation of the S's would be an equivalent solution.

The quantity $\tilde{\theta}$, together with the constraints mentioned in the paragraph above, are not so designed as to ensure that the resulting S's are even monotonic with increase in the k's . Certainly no linear relation with a formula score or number rights is implied nor should it be, since the difficulty levels of the tests are intentionally mismatched. However, it might be desirable to attempt some smoothing of the solution by an averaging process such as is used in the quantity θ below, which fits the total of an examinee's scaled score and the two scaled scores from the adjacent levels on one test to a similar average from the other. The quantity to be minimized with this smoothing is highly similar to $\tilde{\theta}$ except that the range of k is from the

second score level to the next to the last score level; i.e., for test j there is a 0th score level and a $(K_j + 1)$th score level. Then we have

$$\theta = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{j'=1}^{J} \sum_{k=1}^{K_j} \sum_{k'=1}^{K_{j'}} \delta_{ijk}\delta_{ij'k'} [S_{j(k-1)} + S_{jk} + S_{j(k+1)} - S_{j'(k'-1)} - S_{j'k'} - S_{j'(k'+1)}]^2$$

The reader will note that the deltas actually are the data of this scheme and should be aware that they are in part determined by the design for the collection of data and in part determined by the test performances. The class of adequate designs for a scaling study may be investigated by seeking configurations of deltas that lead to unique solutions for the S's up to a linear transformation. It is not the intent of this paper to go into such designs as the present paper stops by choice with the suggestion of plausible objective functions. It suffices to comment that designs and calculations can be worked out which lead to satisfactory minima of θ or $\tilde{\theta}$. The design suggested in the paragraph above is a quite satisfactory one.

Equating with Mixed Essay and Objective Items

The approach to equating suggested previously can also be applied to the situation where mixed essay and objective tests are used. Despite considerable evidence to the contrary which is convincing to the author, it is often deemed desirable to mix essay material with objective material when testing for achievement evaluation in certain subject matters. Many feel that essay testing is useful in a teaching context and some feel that the utility would carry over in mass achievement testing. And judging from history, the inclusion of essay material will continue at least from time to time within the foreseeable future,

for proponents of essay testing feel that the requirement to write on the test might engender a felt need to learn to write on the part of the candidate. Therefore, despite the author's belief that essay testing in large commercial programs is at best an expensive method of unsuccessfully testing the ability to organize and think, it seems that an equating method will be needed, and one is herein proposed that is consistent with the philosophy of the previous discussion.

It is assumed that each examinee will answer some number of essay questions but that they need not all answer the same questions. Each candidate also takes a section of objective questions within which is embedded a miniature test which has appeared embedded in another objective section for which a reporting score scale is established. In fact, it will be assumed that there are two score scales, one which is finely graduated and one which has rather gross divisions. Scores on the finely graduated scale will be referred to as being on the F scale. The gross scale is given as a series of ranges, or cut points, on the F scale, thus, if equating is accomplished on the F scale, it can be immediately translated into the gross scale and the gross scale need not be considered further here. This gross scale is mentioned because its use may be consistent with the College Board Advanced Placement (AP) practice of reporting scores on only five levels, but reporting on that scale does not require or even recommend equating on that scale (as it is not at this time). Conversion to the gross scale can be done as a last step in the score reporting process, for instance, by establishing or having established a set of cut points t_g , $g = 1...k$ where k is one less than the number of intervals to be reported. If $T_{\tilde{g}}$, $\tilde{g} = 0...k$, is the reported score and if $F \leq t_1$, then T_0 is reported. Otherwise if $t_{g'}$ is the largest of the set of t's which are equal to or less than F , then $T_{g'}$ is reported. It is consistent with some practices to take $T_g = g + 1$.

The sense in which equating will be accomplished is that equal scores on the equating section will imply equal reported scores so far as is feasible within other constraints. Rather than using the actual reported scores, though, the equating will be done using the F scale as indicated above. Prior to the actual equating steps it is necessary to infer for each student a score on the essay section. This score will then be combined with the score on the objective section in an equating step.

In inferring an examinee's score on the essay section it will be recognized in the notation to follow that each combination of reader, question, and examinee is unique in a sense. That is, in the symbolism to follow there is an indexing subscript which refers to reader-question combination and does not abstract quantities related in theory to reader alone, or question alone, though some comments about reader evaluation will follow later in the text. It should also be noted that a symbol W is used to indicate whether or not a particular reader-question combination evaluated a response by a particular student. This quantity takes on values zero and unity, indicating whether or not an observation was generated by a combination of student, question, and reader. The quantities W could be regulated by experimental design, at least as far as readers and questions go and possibly somewhat for students also, though it is assumed that the pattern of W's is left mostly to happenstance. Let

i be a subscript for candidate;

j refer to reader-question combination;

S_i be an unequated score for the ith candidate's essays;

A_j , B_j be constants associated with reader-question combination j which account for the variability and toughness (of the question or the reader) of the particular reader when reading the particular question;

X_{ij} be a numerical score assigned to student i by the jth reader-question combination;

W_{ij} be unity if candidate i received a score from reader-question combination j, zero otherwise; and

C_j be an arbitrary weight chosen to vary the emphasis placed on a particular reader-question combination.

Essentially a single factor system is assumed[8] where the A's and B's adjust for severity of grading and variation in grades assigned, but the same student score underlies all performances by that student. This assumes that scores that are used are not subject to a sliding standard as the readings progress and that careful reader training has taken place so that readers have a common and reliable notion of that which is to be graded. We chose as an objective function

$$\sum_j C_j \sum_i W_{ij} (X_{ij} - A_j - B_j S_i)^2 ,$$

to be minimized subject to the constraints that $\sum_i S_i = 0$ and $\sum_i S_i^2 = 1$. Note that in the expression above if a particular set of A's, B's, and S's are a solution, then adding a constant to the S's and at the same time subtracting from the A's that constant times the B's produces an equivalent solution. Also, the effect of multiplying the S's by a constant is offset by dividing the B's by the same constant. That is, the S's are determined only up to a linear transformation (that transformation is to be determined in connection with the equating). Adding a constant for reader-question combination

---

[8] This assumption is not entirely inconsistent with the conclusion of Torgerson and Green (1952) who, though they noted four factors in the grades assigned by English essay readers, found that a large general factor was dominant.

(j) can be offset by subtracting that constant from the A's for that subscript (j). Further, if all the X's are multiplied by a constant, a solution can be retained by multiplying the B's by the constant. Hence the solution adjusts a tendency toward additive bias (toughness), and the scale of grading may be expanded or contracted arbitrarily if the gradations are not modified. However, a limited use of the grading scale lessens a reader's effect in the objective function. One might try to offset this effect by choosing different values for the C's, or by standardizing a particular reader's grades prior to the minimization of the objective function.[9] If standardization is done and a reader tends to get small values for B, then he probably is not grading on the same standard as the others and some judgment about the suitability of his judgments would be needed.

Minimization of the objective function above requires certain characteristics of the quantities involved. First, to use the scoring procedure $W_{.j}$ must not equal zero (reader-question combination must assign a grade). In fact, if S's were known, it can be seen that two observations are required to allow a unique solution for A and B. Second, the number of constants to be determined less the number of constraints is $2J + I - 2$, where I and J are the number of candidates and the number of reader-question combinations, respectively, and hence $\Sigma\Sigma_{ij} w_{ij}$ must exceed that number. Third, it should not be possible to order reader-question combinations and examinees so that the objective function partitions into more than one sum containing different parameters. That is, if the objective function could be written as $Q_1 + Q_2$ where none of the parameters in Q_1 are contained in Q_2 and vice versa, then separate constraints would be

---

[9] The C's may stimulate more confusion than insight and their inclusion may have been in error. However, they do represent a way of influencing the solution which might at a later point in time prove to have an unanticipated advantage.

required for the S's in each Q . This would occur, for example, if students were allowed to answer only one question, and if this were done, the cases contributing to Q_1 could be equated separately from those contributing to Q_2 though this would probably not be desirable. Further, it should also be quite clear that for each examinee, at least one observation is required from a reader-question combination which has a nonzero weight. Finally, the actual computations can be accomplished using existing logic which has been described by Tompkins (1968).

To complete the description of the equating let

R_y be the average reported score in the last administration for those who received a score of y on the equating test;

S_y be the average essay score (determined using the method above) for those who received a score of y on the equating test; and

U_y be the average score on the multiple choice section for those who received a score of y on the equating test. The multiple choice section referred to here is the current one, not the one entering into R_y .

Then find m , p , and q such that $m S_y + p U_y + q$ is the weighted least squares fit to R_y , where the weights are frequencies associated with equating test score in the administration where the equating is taking place. The conversion to the F scale is $m S_i + p U_i + q$ where the subscripts now refer to examinees rather than to equating test levels.

In this type of equating system the relative weight given to the essay and objective score either passes out of the hands of examiners once the system is implemented or is undefined, or both depending on how one views it. It passes out of their hands once the equating system is in effect because the equating determines the weighting system. If the examiners decide on relative weights when the system is started, feel that these weights are effective, do not change test content, and the candidate populations do not differ drastically, then their weighting is preserved. This weighting is undefined, however, in the sense that common factor variance is not uniquely ascribable to the essay or the objective sections and hence the relative weight is more or less undefined (this is true of all section weightings at the present time as far as the author knows). Also, the t's are fixed in a sense once such a system begins operation. Hence, in using this type of system, a good bit of judgment passes from the operators of the system to the statistical system -- a development which may be viewed with different emotions by different people and mixed emotions by some.

Scaling

As a topic in the psychometric literature, scaling is quite general dealing with systems of assigning numbers to events. One defines a set of equivalent events and then sets up an equivalence between these events and the number system. For example, in the equating system just discussed, the equivalent events were the scores on the equating tests, and the number system being chosen is to be found by minimizing the objective function, perhaps with some apparently necessary constraints. However, local usage assigns the term "scaling" to the alignment of test scores and distinguishes between scaling and equating in terms of the kinds of events chosen as equivalent. Specifically, when the

equivalent events are defined in terms of scores on a performance of a sort not measuring the performance for which the number system is being developed, then local usage has it that this is "scaling." In the present context this latter usage of the word will apply.

Scaling is useful in the applications of measurement systems that use tests in that it is the means by which alignment is achieved among score scales based on variables which measure different things. For example, the Army classification Battery consists of a number of different measures such as clerical speed, numerical ability, mechanical knowlege, etc. However, the scores on these tests are in theory aligned in such a way that they would all have the same mean (one hundred), and standard deviation (twenty), in a World War II mobilization population. Thus when a score of 150 is encountered, it is a good one no matter what the content of the test. Thus some intuitive feel for the size of the scores is established, and the process of explaining the meaning of test scores is quite a bit simpler than if such standardization were not effected.

Another use of scores from a battery, not one recommended by the writer but nevertheless one that is said to occur, is that of using the average of available scores as an argument in a regression function. For example, if a student offers scores on English and Spanish achievement and another offers scores on math and history achievement, the averages of the scores might be substituted into a regression function as "achievement averages." Granted that such use is quite inappropriate, the damage done would be minimized to the extent that some standardization between the achievement tests has occurred. The really appropriate standardization for this purpose is a validity scaling, but where such is not possible, an aptitude-oriented scaling is commonly used. It is the aptitude-oriented scaling, or more generally, situations formally similar to that of aptitude scaling that is the topic of this discussion.

First, consider the problem of using performance data from several sending institutions (secondary schools or undergraduate institutions) to predict performance data at a single receiving institution. Here there is a single criterion hence the predictors should be put on a single appropriate scale. It has been mentioned above that the appropriate scale is the criterion scale in this case, obviously, and methods are available for such scaling (Novick, Jackson, Thayer, & Cole, 1971; Tucker, 1960). However, it has been the practice often to use a test to put the grades on a scale using the Pearson assumptions. If one chooses to go along with the aptitude scaling but realizes the probable incorrectness of the Pearson assumptions, one might proceed by defining a reference population in terms of standard frequencies associated with the levels of the test used for scaling and then transform the grades by equating means and standard deviations. In this way, one works toward the same intent as current methods, that is to produce a scaling that yields similar distributions of the scaling variable. The following notation will be needed. Let

x refer to levels of the aptitude scaling variable,

i refer to the sending institution,

j refer to the particular person within sending institution and at score level x,

y_{ixj} be the grade of the jth person at institution i who received an aptitude score at the xth level,

n_{ix} be the number of people at institution i who received an aptitude score at the xth, level,

A_i, B_i be the multiplicative and additive constants in the transformation at institution i,

$$\bar{y}_{ix.} = (1/n_{ix}) \sum_j y_{ixj} ,$$

$$s^2_{ix} = (1/n_{ix}) \sum_j (y_{ixj} - \bar{y}_{ix.})^2 .$$

In the notation above and to follow, the dotted subscript indicates summation over the subscript position dotted in accordance with the current common practice. Then if a reference population Π had frequencies f_x, the variance σ_i^2 of y in Π would be

$$\sigma_i^2 = \sum_x f_x s_{ix} + \sum_x f_x (\bar{y}_{ix.} - \bar{y}_{i..})^2 , \tag{18}$$

where

$$\bar{y}_{i..} = (\sum_x f_x \bar{y}_{ix.})/(\sum_x f_x) , \tag{19}$$

and assuming that s^2_{ix} and $\bar{y}_{ix.}$ remain unchanged, that is that selection is explicit on x and therefore Equation (1) applies when r and s are defined as one. Then if in Π the scaled variance should be υ^2 and the mean should be μ, then by the usual formulae for equating means and variances

$$A_i = \upsilon/\sigma_i \qquad \text{and} \qquad B_i = \mu - A_i \bar{y}_{i..} .$$

The reader will note that here the computations are given where in other sections they are not. This is for two reasons. First, there is no objective function defining this scaling method. The equalities which are sought can be matched exactly with the data at hand in every case, provided the f's are suitably defined. Second, the f's in Equations (18) and (19) are a source of difficulty which the equations exhibit. Note that if no cases are observed at a level of x at a sending institution, the computations in (18) and (19) are not defined unless f is nonzero only for those levels of x for which n_{ix} is nonzero

for all i. In words, this means that a score level to be used in the scaling must have cases at all sending institutions.

Consider the case where institutions of nonoverlapping quality in terms of aptitude are to be included in a scaling using the above method. Then the scaling can't be performed since overlap must exist; if overlap does not exist, what does scaling with the aptitude test mean? This question is not idle because, for one thing, if <u>any</u> two sending institutions are nonoverlapping the difficulty arises, and for another, the scaling is used in undergraduate and graduate admissions testing programs where the scaling is done on very disparate groups. To clarify the shift from scaling grades to scaling in admissions testing programs, note that the levels of x defined as levels of a single aptitude variable in the case of scaling grades can be taken as categories defined by pairs of aptitude test scores as in the Graduate Record Examinations and the College Board Admissions Testing Program. Rather than indexing the sending institution, the i can be taken as the particular achievement area. Then it can be seen that all the notation leading to the computations of Equations (18) and (19) go through (with a suitable redefinition of υ and μ, and if y is taken as an achievement test score to be scaled rather than a grade point average). Both of these testing programs have achievement areas that serve very able candidates for the most part and some that serve examinees that are relatively much less able. Thus one might expect that quite a number of combinations of aptitude levels do not occur at nonzero frequency for all achievement areas. It might be very difficult to define a meaningful (to substantively oriented users) reference population such that f_x is nonzero for all levels of x for all achievement

populations and yet that seems to be required once the Pearson assumptions are foregone. One is therefore tempted to accept the Pearson assumptions anyway, but it should be pointed out that it is precisely where the various populations to be scaled are disparate that the problem with the Pearson assumptions is intensified.

The reader might wonder why it is insisted that a given level of x must be observed in <u>all</u> populations to be observed. Why not just scale the achievement tests one at a time with the aptitude tests using a definition of f_x that yields a suitable mean and standard deviation of the aptitude test scores? The answer is in the question--the scaling in that case would be to the aptitude test and not particularly to the other achievement tests for whom comparability is claimed. One could assert that this achieves comparability among achievement tests and probably many credulous users would believe it. Their credulity would probably lead them to ascribe a transitivity to the comparability, a belief that is unfounded in logic or fact.

The foregoing discussion of scaling is admittedly quite unsatisfactory to the reader who is looking for definitive answers. All that is said is that what has been done in the past is doubtful, and alternative suggestions are admittedly only partially satisfactory. Where criterion scaling is feasible, it seems to the author that it should be done. But the bases for requiring comparability among types of measures that gain value as a collection because of their differentness need development. It may be that the problems that one attempts to solve through scaling would after careful consideration prove to require another kind of solution altogether.

References

Angoff, W. H. Basic equations in scaling and equating, Appendix IX. In S. S. Wilkes, Scaling and equating College Board tests. Princeton, N. J.: Educational Testing Service, 1961. (Also Statistical Report 61-51. Princeton, N. J.: Educational Testing Service, 1961.)

Angoff, W H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D. C.: American Council on Education, 1971.

Boldt, R. F. Study of linearity and homoscedasticity of test scores in the chance range. Educational and Psychological Measurement, 1968, 28, 47-60. (Also College Board Research Development Report 66-7, No. 6, and ETS Research Bulletin 66-43. Princeton, N. J.: Educational Testing Service, 1966.)

Boldt, R. F. Comparability of different tests on the same scale. Research Bulletin 71-10. Princeton, N. J.: Educational Testing Service, 1971.

Boldt, R. F. The inverted-student density and test scores. Research Bulletin. Princeton, N. J.: Educational Testing Service, 1972, in press.

Brogden, H. E. An approach to the problem of differential prediction. Psychometrika, 1946, 11, 139-154.

Brogden, H. E. A simple proof of a personnel classification theorem. Psychometrika, 1954, 18, 205-208.

Brogden, H. E. Least squares estimates and optimal classification. Psychometrika, 1955, 20, 249-252.

Federer, W. T. Procedures and designs useful for screening material in selection and allocation, with a bibliography. Biometrika, 1963, 19, 553-587.

Fiacco, A., & McCormick, G. Nonlinear programming: Sequential unconstrained minimization techniques. New York: Wiley, 1968.

Flaugher, R. L., & Rock, D. A. The wide range validity of certain new aptitude tests. Research Bulletin 66-54. Princeton, N. J.: Educational Testing Service, 1966.

Flaugher, R. L., & Rock, D. A. A fixed length optimal test battery for colleges characterized by diverse levels. AERA Journal, 1968, 5, 659-674.

Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.

Johnson, C. D., & Sorenson, R. C. Model sampling experimentation for manpower planning. Paper given at NATO Conference on Manpower Planning Models, 1971.

Keats, J. A. Estimation of error variances of test scores. Psychometrika, 1957, 22, 29-41.

Kendall, M. G. The advance theory of statistics. London: Griffin, 1948.

Lawley, D. A note on Karl Pearson's selection formulae. Royal Society of Edinburgh, Proceedings, Section A, 1943-4, 62, 28, 30.

Levine, R. S. Equating the score scales of alternate forms administered to samples of different ability. Research Bulletin 55-23. Princeton, N. J.: Educational Testing Service, 1955.

Lord, F. M. Equating test scores--a maximum likelihood solution. Psychometrika, 1955, 20, 193-200.

Lord, F. M. A strong true-score theory, with applications. Psychometrika, 1965, 30, 239-269.

Lord, F. M. Interoffice memorandum, 1969.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. (With contributions by A. Birnbaum.) Reading, Mass.: Addison-Wesley, 1968.

Marshall, A. W., & Olkin, I. A general approach to some screening and classification problems. Journal of the Royal Statistical Society, 1968, 30, 407-443.

McGee, V. E. Towards a maximally efficient system of braiding for Scholastic Aptitude Test equating, Appendix VII. In S. S. Wilkes, Scaling and equating College Board tests. Princeton, N. J.: Educational Testing Service, 1961.

Meredith, W. Some results based on a general stochastic model for mental tests. Psychometrika, 1965, 30, 419-440.

Mollenkopf, W. G. Variation of the standard error of measurement. Psychometrika, 1949, 14, 189-229.

Novick, M. R., Jackson, P. H., Thayer, D. T., & Cole, N. S. Applications of Bayesian methods to the prediction of academic performance. Research Bulletin 71-18. Princeton, N. J.: Educational Testing Service, 1971.

Novick, M. R., & Thayer, D. T. An investigation of the accuracy of the Pearson selection formulas. Research Memorandum 69-22. Princeton, N. J.: Educational Testing Service, 1969.

Pearson, K. Mathematical contributions to the theory of evolution. XI. On the influence of natural selection on the variability and correlation of organs. Philosophical Transactions of the Royal Society of London, Series A, 1903, 200, 1-66.

Pearson, K. On the influence of double selection on the variation and correlation of two characters. Biometrika, 1908, 6, 111-112.

Pearson, K. On the general theory of the influence of selection on correlation and variation. Biometrika, 1912, 8, 437-443.

Press, S. J. Applied multivariate analysis. New York: Holt, Rinehart, & Winston, 1972.

Raiffa,H., & Schlaifer, R. Applied statistical decision theory. Boston: Harvard University, 1961.

Tompkins, L. A factor analytic model. Statistical Systems Report 1. Princeton, N. J.: Educational Testing Service, 1968.

Torgerson, W. S., & Green, B. F. A factor analysis of English essay readers. Journal of Educational Psychology, 1952, 43, 354-363.

Tucker, L. R. Formal models for a central prediction system. Research Bulletin 60-14. Princeton, N. J.: Educational Testing Service, 1960.

APPENDIX

True Score Selection Model

It has been commented that the author has found no way to get around the assumption that explicit selection is on the equating test when the populations tested in an equating experiment using an equating test are not equivalent. This is not quite true, in that the method due to Levine (1955) follows if selection is on the true score. However, that method assumes homoscedasticity with variation in the true score range, an assumption which is highly questionable. It is unfortunate that the assumptions are questionable because the model could be applied in a way that has advantages that will be described in the latter portion of this Appendix.

The following develops an equating method using the notion of true score selection. Let

i be a subscript for operational test;

j be a subscript for equating test;

k be a subscript for candidate (nested within ij);

y_{ijk} be the score of the ith operational test of the kth candidate who took the jth equating test at the ith administration;

x_{ijk} be the score on the jth equating test of the kth candidate who took the jth equating test at the ith administration;

δ_{ij} be 1 if equating test j was taken with operational test i, zero otherwise;

t_{ijk} be the true score of the kth candidate taking the jth equating test at the ith administration; and

ρ_i be the reliability of the ith operational test.

Assume: $y_{ijk} = a_i + b_i t_{ijk} + e_{ijk}$

$x_{ijk} = c_j + d_j t_{ijk} + \varepsilon_{ijk}$

$\bar{t}_{i..} = \bar{t}_{ij.}$

$\bar{e}_{ij.} = \bar{\varepsilon}_{ij.} = 0$

$Cov_{te} = Cov_{t\varepsilon} = Cov_{e\varepsilon} = 0$

$_{ij}\sigma_t^2 = {}_i\sigma_t^2 \quad {}_{ij}\sigma_e^2 = {}_i\sigma_e^2 \quad {}_{ij}\sigma_\varepsilon^2 = {}_i\sigma_\varepsilon^2$.

The above assumptions are heavily influenced by a belief in the equivalence of the populations which take the various experimental sections of the SAT; i.e., the adequacy of the spiraling operation is accepted. Then

$$_{ij}Cov_{xy} = b_i d_j \, {}_i\sigma_t^2$$

$$_i\sigma_y^2 = b_i^2 \, {}_i\sigma_t^2 + {}_i\sigma_e^2$$

and $$\rho_i \, {}_i\sigma_y^2 = b_i^2 \, {}_i\sigma_t^2 \ .$$

The deductions immediately above give some observables on the left in terms of structural variables on the right. In these equations the reliability is assumed to be observed as the result of split-half scoring of the operational test, correlating the half-length tests and then correcting to full-length. Such a method is feasible using modern computing equipment and is preferable to other approximations (since the approximation is not necessary, and may very well not agree with the preferable split-half method).

Note that the covariance ($_{ij}Cov_{xy}$) between operational and equating test can in theory be split into two multiplicative parts, one associated with the

operational test and the other associated with the equating test. This could be accomplished in a least squares way as follows. Find $\{\sigma_i\}$, $\{\beta_j\}$ such that

$$\sum_i \sum_j \delta_{ij} ({}_{ij}\mathrm{Cov}_{xy} - \alpha_i \beta_j)^2 \Big|_{\min}$$

is at a minimum with arbitrary norming on the α's or β's . Take

$$\alpha_i = K b_{i\,i} \sigma_t^2$$

and

$$\beta_j = d_j / K$$

where K is an arbitrary norming constant. Then

$$(\rho_{i\,i} \sigma_y^2)/(\alpha_i) = b_i / K \ .$$

Find $\{C_j\}$, $\{G_i\}$ such that

$$\sum_i \sum_j \delta_{ij} (\bar{X}_{ij.} - c_j - \beta_j G_i)^2$$

is at a minimum subject to arbitrary additive norming on the G's . Then

$$G_i = K \bar{T}_{i..} + P$$

where P is an arbitrary norming constant. Then

$$\bar{Y}_{i..} = a_i + \frac{\rho_{i\,i} \sigma_y^2}{\alpha_i} K \bar{T}_{i..} = a_i + \frac{\rho_{i\,i} \sigma_{iy}^2}{\alpha_i} G_i - \frac{\rho_{i\,i} \sigma_{iy}^2}{\alpha_i} P$$

or

$$a_i = \bar{Y}_{i..} - \frac{\rho_{i\,i}\sigma^2_y}{\alpha_i}(G_i - P) .$$

In terms of the various quantities developed, then,

$$y_{ijk} = \bar{y}_{i..} - \frac{\rho_{i\,i}\sigma^2_y}{\alpha_i}(G_i - P) + \frac{\rho_{i\,i}\sigma^2_y}{\alpha_i} Kt_{ijk} + e_{ijk} .$$

To equate one would choose constants A_i and B_i such that if the true scores on some agreed upon scale and the reporting score scale is such that there exist constants γ and η,

$$E(S) = \gamma_t + \eta ,$$

where $S_i = A_i y_{ijk} + D_i$, then coefficients of powers of t can be equated to γ and η if P and K are known, and A_i and B_i can be easily solved for. To find P and K, arbitrary scaling can be put on the true scores. For example, it might be desired that for some combinations of administrations the average true score should be considered M. Then averaging the G_i's over these administrations one would obtain

$$\bar{G} = KM + P$$

and

$$P = \bar{G} - KM .$$

Thus P could be eliminated, and it should be noted that the average involved could be a weighted one involving numbers of cases at the administrations or some other basis for choosing weights. The quantity K might be chosen so

that the b's are, on the average, equal to one and hence that the product of the reliability coefficient and the test variance would be an estimate of the true score variance for the particular test involved or at least on a scale of the same order of magnitude.

For an equating method, assuming the scale is set so that α's and c's are known, then find B's so that

$$\sum_j ({}_{ij}\mathrm{Cov}_{xy} - \alpha_i\beta_j)^2 \Big|_{\min}$$

is at a minimum and set

$$G_i = (\bar{x}_{ij.} - c_j)/\beta_j \ .$$

Then

$$A_i = \gamma\alpha_i/\rho_{i\ i}\sigma_y^2$$

and

$$B_i = \eta - A_i(\bar{y}_{i..} - \frac{\rho_{i\ i}\sigma_y^2}{\alpha_i} G_i)$$

are the new conversion parameters. The advantages of this method are that only α's and c's are needed for old forms and equating tests. The method is drift free in the sense that one uses all the available data to arrive at weightings which are calculated to relate reported scores to true scores. Of course, variations in the quality of the examinations will affect the result of this type of scaling as it would in any method. Also, minor changes in the method would allow its use with internal equating tests. The important point is that the reliability used should be on material which is not part of the equating

test because if it is, it may be necessary to solve a quadratic equation to obtain a solution and that quadratic equation may not have real roots for some unfortunate administration.

A major weakness of the method is that it suggests to the author no particularly good way to proceed when the end-point problem occurs as it will quite often. Another weakness is that the validity of the model is doubtful, though on this score it is at least as good as the models currently in use. Certainly it is the only model that makes a plausible assumption about the effects of selection of populations.